Deliverable Number: D8.3

Deliverable Title: Database of occupational titles, with explanatory note

Work Package: WP8: - A coding module for socio-economic survey questions

Deliverable type: Other

Dissemination status: Public

Submitted by: University of Amsterdam

Authors: Kea Tijdens (AIAS, University of Amsterdam)

Date Submitted: January 2019
SERISS (Synergies for Europe’s Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe’s social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe’s grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four-year project (2015-'19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey of Health Ageing and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: Tijdens KG (2019) Database of occupational titles, with explanatory note. Deliverable 8.3 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables
# Table of content

## Summary

1. **Introducing SERISS and Deliverable D8.3**
   - Outline of deliverable D8.3
   - The choice of 99 countries and their languages
   - Acknowledgements

2. **The history of the WagelIndicator database of occupations**
   - The start in 2000
   - Contributions from the WOLIWEB and EurOccupations projects
   - Contributions from the search tree and semantic matching
   - Contributions from the InGRID and SERISS projects
   - Contributions from the SSHOC project
   - Researching respondents’ use of the database

3. **The state-of-the-art of occupation survey questions and coding**
   - Introduction
   - Introducing the ISCO Occupational classification
   - The occupation question in surveys
   - A closed occupation question in web-surveys
   - Classifying occupations in a multi-country approach
   - Populating the source list
   - ISCO-08 'not elsewhere classified' occupations
   - Translating the source list
   - Do translated occupational titles have the same job content across countries?
   - Coding the source list

4. **Design requirements for an occupation database aiming at self-selection**
   - Introduction
   - Highly aggregated job titles: clerks, operators or teachers
   - Abbreviations and organisation specific job titles
   - Synonyms
   - Skill levels, professionals, associate professionals and certified occupations
   - Occupations in the corporate hierarchy
   - Composite jobs
   - Handicraft workers and machine-operators
   - Subsistence farmers, fishers, hunters and gatherers
   - Male versus female titles
   - “My job title is not in your list”
   - Start year and end year

5. **The occupation to industry prediction**
   - Developing the occupation to industry prediction
   - The occupation to industry mapping table

6. **The excel file of the occupation database**
   - Introduction
   - Sheet CODESET
   - Sheet SEARCH TREE
   - Sheet MAPPING
   - Sheet LABELSET
   - What the database cannot do

7. **Surveycodings.org: select country and browse the Database live search**
   - Source list
   - Search tree

## References

## Appendix: Instructions to translators

- Introduction
- Translation instructions
Table of Tables

Table 1  The main concepts used in this section ................................................................. 9
Table 2  Code, source list and two look-up tables for two occupational titles ...................... 17
Table 3  Code, source list and two look-up tables for six occupational titles .......................... 18
Table 4  Examples of first line supervisors and department managers ............................... 18
Table 5  Weavers and handicrafts workers ......................................................................... 19
Table 6  Male and female occupational titles for Czech and German ...................................... 20
Table 7  Male and female title in English locales .................................................................. 20

Table of Screenshots

Screenshot 1  Screenshots showing the match list for the highly aggregated occupations teacher and clerk .................................................................................................................................................. 15
Screenshot 2  Screenshot showing that an abbreviation (GP) does not provide a match ........ 16
Screenshot 3  Screenshot showing that a slang word does not provide a match .................... 16
Screenshot 4  Screenshot of the sheet CODESET ................................................................ 24
Screenshot 5  Screenshot of the sheet SEARCH TREE ..................................................... 24
Screenshot 6  Screenshot of the sheet MAPPING .................................................................. 25
Screenshot 7  Screenshot of the sheet LABELSET ............................................................. 25
Screenshot 8  Screenshot of the sheet OCC>>INDUSTRY .................................................. 26
Screenshot 9  Screenshot of the live search of the WISCO occupation database ................. 27
Screenshot 10 Screen shot of the autocomplete box in live search of the WISCO occupation database .............................................................................................................................................. 27
Screenshot 11 Screenshot of the information of the WISCO occupation database .................. 27
Screenshot 12 The search tree in the WISCO occupation database ..................................... 28
Summary

Many questionnaires have a question “What is your occupation?”. The answer is commonly asked as an open text field, challenging the survey-holder to code the response into an occupation classification. Alternatively, in web-surveys respondents can self-identify their occupation from a database. Task 8.2 in SERISS includes five deliverables, D8.3 to D8.8, all contributing to the aim to build a database of occupations for multi-country, multilingual surveys. Departing from the existing WageIndicator occupation database of some 1 700 titles for 35 languages, the database was supplemented with information from coding indexes, and other lists of occupations. The final result was a database with a source list of 4 233 titles, all coded according to ILO’s international ISCO-08 classification. Untranslated entries in the source list were translated by a professional translation agency, resulting in a database serving 47 languages and covering 99 countries.

The database can be used for respondents’ self-identification of occupation in web-surveys. Computer-assisted interviews can be supported too.

The database also includes a tool for the “In which industry are you working?”, providing the most likely industries, given the respondents’ occupations.

The accompanying database occupations_ISCO_for_API_v46_4000_2019-02-04_SERISS_del can be downloaded, see http://surveycodings.org/.
1. Introducing SERISS and Deliverable D8.3

Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) is a four-year project that aims to strengthen and harmonise social science research across Europe (2015-’19). Work Package 8 (WP8) of SERISS aims to provide cross-country harmonised, fast, high-quality and cost-effective coding of open-ended questions on respondents’ occupations, industries and education into international standardized classification systems, and to develop a tool to collect standardized social network information, as described in SERISS Annex 1 (European Commission, 2015). Occupation, industry, employment status, educational attainment and field of education are core variables in many socio-economic and health surveys. In addition, the size and intensity of social networks are key variables in social surveys. However, their measurement, especially in a cross-cultural, cross-national and longitudinal context, is cumbersome, not sufficiently standardized and often expensive. This work package takes recent scientific and technological developments as an opportunity to improve this situation for the benefit of survey measurement quality and to provide cost-effective solutions to Research Infrastructures by developing a survey module with the related survey questions and the databases for answering these questions.

This report concerns Task 8.2 “Compile the API-database of occupations” of WP8.1. The responsible partner is the University of Amsterdam (UvA); partners are SHARE (UNIVE) and SHARE (CentERdata). Task 8.2 aims to compile a database of occupational titles for 99 countries with in total 47 languages, with all occupational titles coded according to the International Standard Classifications of Occupations 2008, abbreviated as ISCO-08 (ILO, 2012).

Task 8.2 consists of five deliverables:
- D8.3 Database of occupations + explanatory note
- D8.4 Validation of ISCO-08 codes + explanatory note
- D8.5 Vacancy crawler and additions to database + explanatory note
- D8.6 Job task collector and additions to database + explanatory note
- D8.7 Database of occupations for five languages + explanatory note

For Deliverable D8.4 the team collected as many coding indexes as could be found, provided they used the ISCO-08 classification and included occupational titles beyond 4-digit. The resulting merged validation database of coding indexes had 70,489 records of ISCO-08 5-digit occupational titles and their 4-digit code from 20 sources with 19 different languages. The accompanying database provides the merged database of coding indexes.

Deliverables D8.5 and D8.6 aim to add “additions to database”. At the time of writing the field work for both D8.5 and D8.6 has been completed, but it turned out that for several reasons hardly any occupational titles could be added to the database. The main reason was that the most titles extracted in D8.5 and D8.6 already were present in the database. A second reason was that the job titles used in vacancies were often too specific for the

---

1 The technology of an API is explained in Martens M (2017) API for databases + explanatory note. Deliverable 8.1 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables
2 Tijdens KG, Kaandorp CS (2018) Validation of ISCO-08 codes + explanatory note. Deliverable 8.4 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables
database, for example by an extensive use of the additions ‘junior’ or ‘senior’ whereas these additions have been left out in the database, as will be explained in section 5 of this report.

The work for Deliverable D8.7 is outsourced to the Institute for Employment Research (IER), University of Warwick, UK. The task relates to the translation of the occupation titles in the database, designed in D8.3, into Russian, Mandarin, Arabic, Hindi and Bahasa. In addition, IER conducted a validation check of the ISCO-08 coding.

Deliverable D8.14 provides the phrasing for the survey question “What is your occupation?” and the hint “If your occupation is not in the list, please select the one that comes closest”. The accompanying database provides the translations of these questions for 47 languages.

Outline of deliverable D8.3

This report concerns deliverable D8.3. The objective of this deliverable was to extend the existing, multilingual WageIndicator occupation database to 99 countries and to 5 000 occupational titles, all coded according to ISCO-08. Deliverable D8.3 builds on preparatory work done for the InGRID project. This project facilitated the design for the development of the occupation database, accompanying this deliverable. Based on this preparatory work, the main output of D8.3 is therefore the accompanying database of occupational titles. In Section 5 of Deliverable D8.3, I have therefore largely copied the design principles detailed in Section 4 of my InGRID working paper M21.4 (Tijdens, 2015a).

This report has an accompanying database, which is downloadable from the website www.surveycodings.org, called occupations_ISCO_for_API_v46_4000_2019-02-04_SERISS_del. The content of the database is described in detail in section 6 of this report. The file contains the source list of occupations and their codes, its translations, the structure of the search tree for online browsing of the database, a mapping table of the codes into the ISCO-08 classification codes, and a labelset for all codes included in the database. The database is used in the live-search at surveycodings.org. Survey holders can use the database to prepare an API or search tree for the occupation question in their surveys. The database is called the WageIndicator Database of ISCO Occupations (WISCO).

The choice of 99 countries and their languages

The SERISS WP8 team developed criteria for the selection of the 99 countries. Three criteria were applied, notably the countries’ membership of the European Union or the OECD, both including associate countries; the countries’ membership in the CAMCES education database and the countries’ membership in at least one survey wave of ESS, WVS, EVS, SHARE, G&G, EWCS, EQLS, ISSP, PIAAC 2012, PISA 2012, and WageIndicator. These criteria were applied in combination with the countries’ population.

---

4 Tijdens KG (2018) Database of occupations for five languages + explanatory note. Deliverable 8.7 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables
5 See the accompanying database for Deliverable D8.14 of the SERISS project. Available at: www.seriss.eu/resources/deliverables
6 InGRID stands for Inclusive Growth Infrastructure Diffusion. This Research Infrastructure was funded by EU-FP7 [Contract no. 312691, 2013-17] and was coordinated by HIVA-KU Leuven, Belgium.
7 Tijdens KG (2016) Survey Q&A + explanatory note. Deliverable D8.14 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables
whereby countries with populations less than 100,000 were to be excluded, and so were non-UN-recognized territories.

<table>
<thead>
<tr>
<th></th>
<th>Criteria</th>
<th># countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EU-28 plus EU-ass13, but excluding Faroe Islands and Liechtenstein</td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>OECD plus OECD-ass, not mentioned above</td>
<td>+ 10</td>
</tr>
<tr>
<td>3</td>
<td>Russian Federation, Egypt (largest Arab speaking country), China, India, Indonesia</td>
<td>+ 5</td>
</tr>
<tr>
<td>4</td>
<td>Countries in CAMCES database, not mentioned above, but excl. Northern Ireland</td>
<td>+ 5</td>
</tr>
<tr>
<td>5</td>
<td>If population &gt; 30 mln and membership in surveys &gt;0 (excl WageIndicator member)</td>
<td>+ 18</td>
</tr>
<tr>
<td>6</td>
<td>If population &gt; 10 mln and membership in surveys &gt; 1 (excl WageIndicator member)</td>
<td>+ 6</td>
</tr>
<tr>
<td>7</td>
<td>If population &gt; 5 mln and included in WageIndicator web-survey and having a non-unique language in order to prevent huge translation costs</td>
<td>+ 16</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>99</td>
</tr>
</tbody>
</table>

These 99 countries have in total 47 languages, whereby only the main language(s) of the country were taken into account. We selected the language spoken by the largest population group. For a few cases, we selected more than one language (Belgium, Switzerland). Note that where possible the same language was adapted to national features, e.g. American English and British English or Brazilian Portuguese and Portuguese Portuguese. The list of countries and languages can be found in SERISS D8.14.

**Acknowledgements**

The WISCO database has been developed gradually since the year 2000, starting with a coded occupation list for the Netherlands, used for self-identification in the WageIndicator web-survey on work and wages. The Foundation is a non-profit organization, established 17-9-2003, and dedicated to labour market transparency by providing accurate wage and wage related information, posted on its national websites in almost a hundred countries. Its founders are the University of Amsterdam/AIAS, the Dutch Confederation of Trade Unions (FNV), and career website Monster.

From 2004 onwards, the WageIndicator websites and its web survey was gradually extended to currently 100 countries, and so was the WISCO database. This could be done thanks to the following projects: WOLIWEB in eight EU countries (EU-FP6, no. 506590, 2004-’06), EurOccupations (EU-FP6, no. 028987, 2006-’09), GLOBAL—Towards a Global WageIndicator (FNV Mondiaal-Netherlands, 2008-’10), DECISIONS FOR LIFE in non-European countries (MDG 3 Ministry Foreign Affairs–Netherlands, 2008-’12), WISUTIL for occupations in the energy sector (EU-Social Dialogue programme, no. VS/2010/0382, 2010-’11), WICARE for occupations in the care sector (EU-Social Dialogue programme, no. VS/2013/0404, 2013-’14), InGRID (EU-FP7, no. 312691, 2013-’17), EDUWORKS (Marie Curie Initial Training Network, no. 608311, 2013-’17), SERISS (EU-H2020, no. 654221, 2015-’19), and SSHOC (EU-H2020, no. 823782, 2019-’22). Acknowledgements are further extended to Eurofound for providing the verbatims of the occupation question in its European Working Conditions Survey 2010, to the statistical office of Portugal INE for providing its coding index, and to WageIndicator team members who over the past decade have contributed to the database for their own languages.
2. The history of the WageIndicator database of occupations

The start in 2000
The very first version of the WageIndicator database of occupations dates back to 2000, when the author, in cooperation with colleagues, conducted a survey in the Netherlands, aiming to collect up-to-date information about women’s wages by occupation. Hence, large sample sizes were needed to meet the demands of sufficient observations for not-frequent occupations. A survey was printed in women’s magazines and simultaneously a web-survey was published on a frequently visited website for women, together resulting in almost 14,000 completed questionnaires. To prevent the high costs for office coding of occupational titles, both the printed and the web-survey included a list of the 55 most frequent occupational titles for women, clustered into seven groups with an option ‘other’ with a text box at the bottom of each group. Slightly over one quarter of the respondents used the text box to enter a job title not listed, which were then recoded into 19 additional occupational titles.

In 2001, the survey data were used to start a WageIndicator website with a so-called Salary Check for women, attracting large numbers of web-visitors. The web-survey was posted on this website and visitors were invited to complete the survey, which was now addressing both women and men. For the occupation question, the list of women’s occupations was extended with the most frequent men’s occupations and all were coded according to the Standard Occupational Classification 1996 from Statistics Netherlands.

Contributions from the WOLIWEB and EurOccupations projects
In 2004, as part of the EU-funded WOLIWEB project (EU-FP6, no. 506590), WageIndicator could start websites with continuous web-surveys on work and wages in seven large EU countries. For the occupation survey question, the choice was again in favour of a predefined list with a search tree and not in favour of post-coding text strings, due to budgetary considerations. To populate the list of occupations, project partners were asked to send coding indexes or other lists of occupations from their countries, and provide English translations of the titles. This list of English titles was named the source list. By 2006, the source list contained more than 5,000 occupational titles, many of them present in one other language only. Occupations from different countries with the same English occupational title were merged. For most countries the database included a list between 500 and 1,500 occupations. The aggregation level of the occupational titles varied across the countries and included both very broad and very narrow titles. Where possible, the titles in the source list were coded to ISCO-88, otherwise they received a follow-up number. The search tree in the web-survey was extended from 2 to 3 levels. In 2005 and 2006, some countries added occupational titles to the list, based on feedback from web visitors who could not identify their occupational title. Based on projects with non-EU countries, the occupation list was translated into Portuguese, Russian, Chinese, Indonesian, and a few other languages.

In 2006, the EurOccupations project (EU-FP6, no. 028987) made it possible to improve the database. Based on Draft 3 of ISCO-08, published in September 2006, and on ILO’s alphabetical index of occupational titles for ISCO-88(COM), the source list was revised and coded according ISCO-08, resulting in 1,433 occupational titles, all at 5 digit level of the ISCO classification. Project partners translated the occupational titles for the languages of the eight EU countries in the project. This resulted in a database with hardly any not-translated occupational titles.

In spring 2008, ILO published the final draft ISCO-08 classification with 433 so-called occupational units at 4-digit level. Compared to Draft 3 (ILO, 2006), a number of occupations
were assigned a different skill level.\textsuperscript{8} The codes in the source list of occupations were adapted accordingly. Meanwhile, WageIndicator partners in non-EurOccupations countries had commented on the source list, noticing translation difficulties and adding missing occupational titles. The source list was then critically reviewed with regard to internal consistency, extended where possible, and all entries were located within the 3-level search tree of the web-survey. By early 2009 the source list included 1 594 distinct occupational titles with translations for approximately 30 languages for almost 50 countries in and outside Europe. The team made a cross-over table between the occupation codes used until 2009 and the occupation codes used from 2009 onwards. The database was named the WageIndicator Database of ISCO Occupations (WISCO). The primary aim of this database was to facilitate valid self-identification of occupation in multi-country, multilingual web-surveys by means of a search tree (Tijdens, 2010).

**Contributions from the search tree and semantic matching**

In 2009, the revised WISCO database was uploaded in the WageIndicator web-survey for the - at that time - 50 countries with a WageIndicator website and a web-survey. Data collection started for some countries as of July 2009, but for most countries as of October 2009. The database and the search tree were also implemented in the Salary Checks in the WageIndicator websites. With millions of web visitors and over a hundred thousand respondents per year, the search tree and the occupation database were tested extensively. Over the years WageIndicator has received few complaints that ‘my occupation is not in your list’, leading to the conclusion that most respondents are able to recognise their occupational title and that the database serves the aim of self-identification for large numbers of jobholders.

In the years after 2009, occupations have been added to the database, mostly because web visitors requested so in their emails. For the UK a number of management occupations and for the Czech Republic and Slovakia medical specialists have been added. For some titles translations for other languages were available, for other titles they were not. For Germany the skill levels for some skilled occupations have been further detailed by distinguishing occupations at university and higher vocational level. By 2015, the source list in the database held 1 896 occupational titles, of which 132 were country specific, and thus not translated in other languages. The number of occupational titles in the database varied slightly across countries, because in some countries some occupation titles could not be translated because the words did not exist in the language at stake or because distinct occupational titles in the source list were translated similarly and the duplicates were removed.

In 2012, the search tree in the WageIndicator web-survey and Salary Check was extended with an autosuggest box, hence a Google search type. Such semantic matching allows visitors to self-identify their occupation by typing text whereby matches with words in the list of occupations are instantly shown. Visitors can then select the most relevant match. In the past years, the autosuggest box has become more important in the web-survey, specifically because web-surveys are increasingly completed on smart phones, and on this device search trees do not fit the small screens. Hence, for smart phones the WageIndicator web-survey only uses the autosuggest box.

**Contributions from the InGRID and SERISS projects**

Although a source list of 1 896 occupational titles may seem a large number, one has to take into account that a labour market in any country easily can include 10 000 or more job titles. The use of the database for self-identification in web-surveys would therefore profit from extending the number of occupations as well as the number of languages, to provide a

\textsuperscript{8} The final ISCO-08 coding index was published in 2012 (ILO, 2012)
global occupational database. In 2014, the InGRID project provided funding for the systems design of a global tool for the measurement of occupations in web-surveys, resulting in a paper about the design principles for a database to be used for the measurement of occupations globally (Tijdens, 2015a, 2015b).

In 2015, the SERISS project (EU-H2020, no. 654221) provided funding to populate the database for more entries and more languages, according to the design principles outlined, and to develop an Application Programming Interface (API) for the measurement of occupations in web-surveys. Thanks to SERISS the 1 700 occupational titles and the 35 languages could be extended. The initial plan was to extend the database to 5 000 titles. However, after the database had been extended to more than 4 000 titles, the newly added titles increasingly were duplicates of the titles already in the database. The final result was a database with a source list of 4 233 titles. The initial plan to extend the database to 47 languages was fully reached and the database now covers 99 countries.

**Contributions from the SSHOC project**

Occupational structures are dynamic and the occupational composition of a labour force varies over time and across countries. Due to technological and organisational changes, new and emerging occupations are noticed. Most surveys ask respondents for their current occupation, but some have questions about father’s and mother’s occupation, mostly specified as the occupation when respondent was 14 years of age. Other surveys address elderly people, who are no longer in paid employment, but are asked to report their last occupation before retirement. This approach challenges the identification of time frames for occupational titles. A straddle carrier driver or a web manager did not exist in periods when the straddle carrier or the web was not yet introduced. For quite some occupational titles in the source list a start year can be assigned. For example, in the case of web managers we could assign as start year 1990.

The SSHOC project (EU-H2020, no. 823782) has provided funding for an extension of the database for the measurement of parent’s occupations, and thus to include 20th century occupations. Where possible, start and end dates of the life cycle of an occupation will be assigned. This will be done in cooperation with the database of historical occupations HISCO that includes occupational titles that don’t exist today, because the machinery, equipment or materials used have become outdated. The HISCO scheme is based on the coding of the 1 000 most frequent male and female occupational titles in datasets from eight different countries: Belgium, Britain, Canada, France, Germany, the Netherlands, Norway and Sweden. The occupational data which were employed to develop the scheme spanning the period 1690-1970, but are mostly from the nineteenth century.

**Researching respondents’ use of the database**

A few studies have addressed respondents’ use of the database for the self-identification of occupation in a web survey and the coding indexes used. The results are summarized here.

One study detailed the use of the search tree (Tijdens 2014b). As in other survey modes, most web surveys use an open-ended question for occupation, though the absence of interviewers elicits unidentifiable or aggregated responses. Unlike other modes, web surveys can use a search tree with an occupation database. They are hardly ever used, but this may change due to technical advancements. This article evaluates a three-step search tree with 1,700 occupational titles, used in the 2010 multilingual WageIndicator web survey for UK, Belgium and Netherlands (22,990 observations). Dropout rates are high; in Step 1 due to unemployed respondents judging the question not to be adequate, and in Step 3 due to

---

search tree item length. Median response times are substantial due to search tree item length, dropout in the next step and invalid occupations ticked. Overall the validity of the occupation data is rather good, 1.7-7.5% of the respondents completing the search tree have ticked an invalid occupation.

Another study explored the use of the option ‘other’ in the occupation search tree (Tijdens, 2015). Can self-identification of occupation be applied in web surveys by using a look-up table with coded occupational titles, in contrast to other survey modes where an open format question with office-coding has to be applied? This article is among the first to explore this approach, using a random sampled web survey (N=3,224) with a three-level search tree with 1,603 occupations and offering a text box at the bottom of each 3rd level list. 67% of respondents ticked in total 585 occupations, of which 349 by at least two respondents and 236 by only one, pointing to a long tail in the distribution. The text box was used by 32% of respondents, adding 207 occupational titles. Multivariate analysis shows that text box use was related to poor search paths and absent occupations. Search paths for five of the 23 first-level entries should be improved and the look-up table should be extended to 3,000 occupations. In this way, text box use and thus expensive manual coding could be reduced substantially. For such large look-up tables semantic matching tools are preferred over search trees to ease respondent’s self-identification and thus self-coding.

Finally, one study explored the codes used in coding indexes (Tijdens and Kaandorp, 2018). Occupational coding in multi-country surveys is mostly a black box: have national survey agencies classified the same occupational titles into the same category across countries? This paper attempts to validate the coding from 5-digit occupational titles into the 4-digit occupational units of the international ISCO-08 classification, based on a comparison of coding indexes from national statistical offices. Two research objectives are central. To what extent are occupational titles in the coding indexes similar, when comparing their English translations? What percentage of similar occupational titles is coded similarly across countries? To answer these questions, we merged titles from 20 coding indexes (18 non-English), resulting in 70,489 records. We translated the titles in English, using online dictionaries and Google translate (4.2% could not be translated). We checked for existent codes of the titles, using ILO’s ISCO-08 coding index (10.3% non-existent). The remaining database had 60,559 records, of which 32% had at least one duplicate title (19,044 records). These duplicate records could be aggregated into 5,350 occupational titles. Only 64% of these titles had the same ISCO-08 4-digit code, 70% at 3-digit, 74% at 2 digit, and 80% at 1-digit. Users of multi-country surveys should be cautious when using the 4-digit ISCO-08 codes.
3. The state-of-the-art of occupation survey questions and coding

Introduction
The section discusses the state-of-the-art of occupational coding and the related survey questions, as well as the aims of the database of occupations. Table 1 presents the main concepts used in this section.

<table>
<thead>
<tr>
<th>Table 1. The main concepts used in this section</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source list</strong></td>
</tr>
<tr>
<td><strong>ISCO</strong></td>
</tr>
<tr>
<td><strong>Code</strong></td>
</tr>
<tr>
<td><strong>Locale</strong></td>
</tr>
<tr>
<td><strong>Translation</strong></td>
</tr>
<tr>
<td><strong>Database</strong></td>
</tr>
<tr>
<td><strong>Look-up table</strong></td>
</tr>
<tr>
<td><strong>Entry</strong></td>
</tr>
</tbody>
</table>

Introducing the ISCO Occupational classification
In 1958, the International Labour Office (ILO) of the United Nations had developed the International Standard of Occupational Classification (ISCO) for the harmonization of the measurement of occupations. Revisions were done in 1968, 1988, and 2008. Today, ISCO has become the standard classification in many countries for their Labour Force Surveys or Censuses. ISCO-08 defines a job as a set of work tasks and duties performed by one person (ILO, 2012). Jobs with the same set of main tasks and duties are aggregated into the so-called 4-digit occupation units. On the basis of similarity in the tasks and duties performed, the units are grouped into 3- and 2-digit groups, which in turn on the basis of the skill level are grouped into 1-digit groups. These groups differentiate between four skill levels, notably unskilled, semi-skilled, skilled and highly skilled, which are related to ISCED, the International Standard Classification of Education.

With its ISCO-08 classification the ILO targets a global use in Labour Force Surveys, to facilitate cross-country comparisons. For this purpose, the ISCO-08 coding index comprises thousands of occupational titles (ILO, 2012). The coding index is in English. Coding indexes in other languages are not available. Following its policy towards an ISCO-08 harmonised occupational classification for European Labour Force Surveys, the European Commission has translated the ISCO-08 4-digit classification into all languages of the European Union to ensure that occupations are coded similarly across countries. Unfortunately, ILO hardly employs any support staff for the maintenance of ISCO and ILO does not have a discussion platform where the classification of new emerging occupations could be debated. This is a pity, specifically when compared to the discussion platform available for the NACE industry classification (EUROSTAT, 2008).

The occupation question in surveys
An inventory of the occupation question in 33 surveys in English, German, French, or Dutch, and held in Europe and the United States showed that 23 were designed for face-to-face interviews, of which some computer-assisted, and the remainders aimed at postal/paper surveys or web-surveys (Tijdens, 2014a). In 25 surveys an open text format for the occupation question was used and almost all included interviewer instructions, such as

---

11 Compare for example to the active SOC revision groups of the Office for National Statistics in the UK.
Avoid vague occupational titles such as manager, clerk, or farmer’ or ‘Write in full details’. The phrasing of the occupation question however varied largely across the surveys. In 6 of the 8 surveys with a closed format question a show card with the ten entries of the ISCO classification was used, which in a few surveys was extended with example occupations within each category. The remaining two surveys provided a show card with a mixture of employment status, occupational titles, skill level, and supervisory position.

Survey-holders usually have manuals to guide interviewers for this survey question. The manual for the US Current Population Survey for example details how interviewers should deal with inadequate descriptions, because these result in difficult to code occupations (US Census Bureau, 2013). Interviewers are instructed that one-word responses to the question on occupation (for example, clerk, engineer, manager, nurse, teacher) are usually far too general to be coded accurately. Whenever very brief responses are given to this question, interviewers should probe to obtain a more specific response. 14 of the 25 surveys with an open-ended question also asked for a job description, for which again the question phrasing largely varied across the surveys. Only a few of the 33 reviewed surveys asked the interviewer to code the occupation during the interview (field-coding), using a show card with the 2-digit occupational units, but most surveys relied on office-coding.

Field-coding is advantageous over office-coding because it allows the interviewer to ask additional information if needed, but in case 4-digit coding is required it needs advanced software on the interviewer’s laptop. Office-coding is recoding at a later point in time and is disadvantageous in budget terms and timelines. Additionally, datasets show that both field- and office-coding may lead to a percentage of unidentifiable occupational titles and a percentage coded at higher aggregation level then the targeted level. A well-known coding software program is CASCOT and its update CASCOT2000.12 CASCOT is among others used by ONS and survey agencies in the United Kingdom and by Statistics Netherlands. In the USA and in Germany auto-coders have been developed (Cheeseman Day, 2014; Bethmann et al., 2014). A huge training set of coded responses is required to apply these machine learning algorithms. A coding tool for office coding is typically designed for high volume batch coding. It is providing a certainty score per coded verbatim. For a semantic matching tool in web-surveys, this functionality is however not relevant. The main purpose of the database of occupations is self-identification by survey respondents by providing each respondent with a list of potential matches when typing his/her occupational title.

To ensure cross-country harmonised coding of the occupation variable extensive cooperation between survey agencies is needed, because multi-country coding indexes do not exist.13 When such cooperation is lacking, the coding of the 5-digit occupational title into the ISCO-08 classification is basically a black box with regard to which extent the same occupational titles are coded similarly across countries. Therefore, a multi-country, multilingual database of occupational titles could meet these requirements, as will be discussed in the remainder of this chapter. Note that the European Commission (2009) has translated the 1-4-digit ISCO-08 coding index in all EU languages, but has not provided translations beyond 4-digits. A recent EU project, called ESCO, has provided translations for almost 3 000 occupational titles. Around 2016 the author has shared the WISCO database with ESCO.

---

12 The Institute for Employment Research (IER), University of Warwick, UK, maintains the coding tool CASCOT, an abbreviation of Computer Assisted Structured Coding Tool. CASCOT is a computer programme designed to make the coding of text information to standard classifications simpler, quicker and more reliable. See http://www2.warwick.ac.uk/fac/sci/ier/software/cascot/internat/, accessed 11-01-2019

13 It may well be that survey agencies have developed their own multi-country coding index, but being commercial undertakings they typically do not share these with a wider public. In countries with a weak NSO and with all surveys done by agencies this may lead to a private index, not a publicly available one.
A closed occupation question in web-surveys

Web-surveys allow for respondent’s self-identification using a look-up database. In web-surveys, three techniques can be used to facilitate searching the database, notably an alphabetically sorted drop-down list, a search tree with two or three levels, and an open format question with semantic matching or autosuggest. All three have been discussed in InGRID working paper M21.4 (Tijdens, 2015). For databases with few entries drop-down lists and search trees can be easily applied, but for large databases the autosuggest option is a better solution. And for web-surveys on smartphones the autosuggest option is better than the search tree because of the small screen size.

Survey respondents tend to report a detailed job title, as they know it from their employment contract, job classification scheme, collective bargaining agreement, job advertisement, or just from a common understanding in the workplace. Therefore, the occupation database cannot consist of ISCO-08 4-digit titles, because these comprise of too aggregated concepts and may lead to so-called aggregation bias. The database should include titles at ISCO-08 5-digit level, which are closest to the detailed job titles of respondents. Some survey respondents, however, tend to report a broad occupational title, such as teacher, clerk, or ‘I work in a shop’. When the database comprises solely 5-digit occupational titles, these respondents are forced to self-identify their occupation in greater detail, because they are presented with multiple options to select from. In summary, the database aims to elicit responses at 5-digit level, which are recoded into the 4-digit ISCO-08 classification and aims to prevent unidentifiable responses and responses at 1-, 2-, or 3-digit level.

Classifying occupations in a multi-country approach

When it comes to classifying 5-digit occupations across countries into ISCO-08 4 digit, two approaches can be distinguished. The first one says that the ILO manual and descriptions are sufficiently detailed and hence it is assumed that national coding of 5-digit occupations leads to similar results across countries. This method is applied in many multi-country surveys, where the field organisations code the occupations for their respective countries. The second approach states that only English occupational titles should be coded, and that therefore national job titles should first be translated and then coded according to the English title. For three countries (Albania, Kosovo & Montenegro) this method is followed for EuroFound’s European Working Conditions Survey 2010 (EWCS): the verbatim responses were translated in English to facilitate central quality control (Gallup Europe, 2010). In retrospect Ganzeboom (2014), applying the first approach in his effort to code parental occupations in the European Social Survey, acknowledges that it would have been much better to ask the coders to translate the occupation titles into English and then code these. Ganzeboom states that Google Translate has become a big help in this respect.

There are costs associated with both approaches. In the first approach the costs are related to the coding by national survey agencies, while no multi-country quality control can be applied. In the second approach translations might be costly, but central coding of the English occupations for the entire multi-country data collection is relatively cheap. In addition, once the translated database has been established it can be reused at very little cost in multiple surveys.

This SERISS Deliverable follows the second approach, departing from an English source list and using translations into other languages. Creating a bespoke database for use in multi-national surveys was preferred over pooling the coding indexes of the national statistical offices. As was shown in SERISS Deliverable 8.4 a pooled coding index included more than 70 000 entries, of which many were not translated in more than one language. This did not contribute to our goal of comparability of occupational titles by means of translations. The results of D8.4 were firstly that a few coding indexes used codes that did not exist in ISCO-08. Of the 20 828 records with duplicate translations in more than one country, only 41% had
the same code for the same occupational title. After aggregating the records into 5,754 occupational titles slightly more than half (54%) had the same code for the same title.

Populating the source list

To populate the source list of a multi-country, multilingual database of occupations, a large number of entries are required to satisfy the large majority of respondents, given that a labour market in any country easily can include 10,000 or more job titles. However, survey respondents will find it difficult to self-identify their occupation from such large databases. To populate the source list of the database of occupations, the following steps have been taken:

• We started from the ISCO-08 coding index with almost 7,000 occupational titles (ILO 2012). This database includes however duplicate titles, which reduces the size of the unique titles by more than half.
• The 1,700 occupational titles of the WageIndicator database were added, thereby removing duplicate titles.
• The comparison of coding indexes of national statistical agencies resulted in new entries to the source list. Only occupational titles that were present in at least three coding indexes were included, after a check for duplicate titles. For the source occupational titles that were present in the database and in the coding indexes, the translations were taken from the coding indexes.
• Given the very skewed distribution of jobholders over occupations, the least frequent titles were ignored, if these occupational titles could be identified. All occupational titles with one jobholder in a country, such as prime minister, king, or alike, were not included.
• We checked the coding agreements across several sources: the codes used in ISCO-08 (ILO 2012), in the WageIndicator database, and in the codes of Statistics Netherlands (CBS 2013). For 181 job titles, this check resulted in different codes for the same job title. The Institute for Employment Research (IER), University of Warwick, UK was asked to provide a final solution for the coding. All titles were then assigned a final code.
• Although the initial plan was to include 5,000 titles in the database, it turned out that increasingly duplicate titles were entered into the database. Therefore, the result was a database with a source list of 4,233 titles.

ISCO-08 ‘not elsewhere classified’ occupations

To ensure high quality coding, for each locale its look-up table of course needs to include at least one 5-digit occupational title in each ISCO-08 4-digit occupational unit. However, in case of office coding unidentifiable occupations are usually classified in the appropriate residual 4-digit occupational category, called ‘not elsewhere classified’, for example Unit Group 1349 ‘Professional Services Managers Not Elsewhere Classified’. ISCO-08 4-digit has 27 residual units. Luckily, ILO (2012) lists examples of occupational titles classified in the residual units. The database includes these occupational titles, and hence all ISCO-08 4-digit units have at least one entry in the source list.

Translating the source list

The following steps have been taken for collecting translations of the occupational titles of the source list:

• The translations of the 1,700 occupational titles of the WageIndicator database were added.

---

14 See SERISS Deliverable D8.4 Validation of ISCO-08 codes
15 Excluding one jobholder occupational titles were also excluded for privacy reasons and to prevent insincere respondents to select that they are a king
• Databases with translations of occupational titles were collected, such as the database of licensed occupations in the European Union, the ESCO occupation database, a database of 300 occupations which were translated from English in several Slavic languages, a file with translated university positions, and databases of occupational titles in Japanese and Arabic and translated into English. These translations were added, provided that the English title was already included in the source list.

• Then the source list plus the available translations were sent out to professional translators, who were asked to translate any occupational titles not yet available in the target language and to check the already translated ones. The translating agency provided two independent translations of each occupational title they translated.

• The translations for five most spoken languages outside the EU28 area, notably Russian, Mandarin, Arabic, Hindi and Bahasa, was outsourced to IER. IER recruited native students in labour market subjects and supervised the translation job.

• We developed an instruction sheet for the translators (see the Appendix of this report) and all translators were aware that occupational titles used in the labour market of the country at stake were preferred over literal translations.

The translations in all languages have been checked for duplicates. If two different source titles were translated similarly, the translated title with the highest skill level was removed. In quite some languages the same two source titles were translated similarly, for example the accountant which is in ISCO group 2, a highly skilled occupation, versus the bookkeeper which is found in ISCO group 4, a semi-skilled occupation. More details about the principles underlying the occupation database and its translations are described in section 5 of this paper. Note that the database should never include two identical titles with different ISCO-08 5-digit codes.

A second translation problem related to non-existent occupational titles. For example, Spain has regional police officers, whereas in the Netherlands such level is not present in the police force. The source list therefore includes this occupation, but it remains untranslated in Dutch for the Netherlands. Of course, the matching list in a Dutch survey will not show the untranslated English source titles.

Finally, we were not able to check country differences for the same languages, with the exception of an extensive check for Portuguese occupational titles in Brazil and Portugal. Quite a few occupational titles from the source list are different in Flanders (Belgium) and the Netherlands and a couple were so in Germany and Austria, but the differences between these countries have not been checked extensively. Any comments regarding the translations are very welcome.

Do translated occupational titles have the same job content across countries?

The translations solely relate to the occupational titles in the source list. ILO’s ISCO-08 classification and coding index assume that occupations with the same titles are classified similarly across countries, otherwise there would not be an argument for a global classification. Hence, the database of occupations challenges whether occupations with the same translated title have the same job content across countries. The answer is that we don’t know whether the national occupational titles with the same title in the source list have the same job content across countries and across languages. We have neither the methods nor

16 We used OneHourTranslation (OHT), a professional translation agency providing translations for requested languages, see https://www.onehourtranslation.com/?cptag=G_Oht_OHT&gclid=EAalqobChMI2-k8uf03vlVA9lyCh0bWwDEEAAYASAEgtdkPD_BwE
17 See SERISS Deliverable D8.7 Database of occupations for five languages + explanatory note
18 Please send comments to the author, k.g.tijdens@uva.nl
the means to test empirically task similarity within occupations across countries and languages. Within the EurOccupations project as well as in the Eduworks and Ingrid projects empirical tests have been undertaken to measure the heterogeneity of tasks within occupations (Tijdens, De Ruijter, De Ruijter, 2012; Tijdens, De Ruijter, De Ruijter, 2014; Visintin et al., 2015). The results show that the frequencies of a given set of tasks within each occupation vary within and across countries, but that these results are limited by sample sizes as very large samples are needed to collect data from a sufficient number of occupations. The tasks-and-duties similarity within occupations within and across countries challenges a body of knowledge that is beyond the measurement of occupations titles.

In an overview of 25 surveys with an open-ended question about occupation, 14 ask for a job description, the so-called tasks-and-duties-question (Tijdens, 2014a). The question is typically asked for office coding purposes, notably when the answer to the occupation question cannot be coded. To the best of my knowledge it is never used to compare job content across countries.

Coding the source list
Each occupational title in the source list of the database has a unique code. The coding rules are as follows:

- position 1-4 = ISCO 08 4 digit code\(^{19}\)
- position 5-6 = follow-up code within ISCO 08 4 digit
- position 7-8 = follow-up code within ISCO 08 4 digit within country specific group, otherwise 00
- position 9-11 = country code for country-specific group, otherwise 000
- position 12-13 = year occupation was added to database since 2017, otherwise 00

Here we discuss an example of the coding. Managerial occupations in farming have code 1311. In the database the occupation ‘Manager livestock farm’ is coded 1311060000000, and ‘Manager horticultural production’ is coded 1311070000000.

---

\(^{19}\) Note that in a text format the codes of the 4-digit military occupations start with a 0, but that in a numeric format the first 0 disappears and therefore the codes of these occupations have three digits instead of four.
4. Design requirements for an occupation database aiming at self-selection

Introduction
The occupations database is primarily intended to facilitate respondents’ self-identification of occupation in web-surveys by means of semantic matching and has been designed accordingly. This section discusses the principles underlying the design of the database and how problems have been solved. In CAPI surveys the database can be used to populated look-up tables to facilitate interviewer field coding or respondent self-identification.

Highly aggregated job titles: clerks, operators or teachers
From office coding it is well-known that some respondents tend to report a highly aggregated job title such as clerk, operator or teacher. These titles are too aggregated to be classified at a 4-digit level. Typically, surveys with open text field provide instructions to prevent this problem. In the database, these highly aggregated occupational titles are therefore not included. If respondents key in ‘teacher’, they are offered a choice of many teaching occupations, and similarly when they key in ‘clerk’, as Screenshot 1 shows. Similarly, no entries for highly aggregated occupational titles such as ‘manager’ or ‘shopkeeper’ are included. The database holds only specific titles.

Screenshot 1 Screenshots showing the match list for the highly aggregated occupations teacher and clerk


Note that this section has largely been copied from Tijdens KG (2015), The design of a tool for the measurement of occupations in web-surveys using a global index of occupations, Working paper, Leuven, InGRID project, M21.4, retrieved from https://inclusivegrowth.be/downloads/output/m21-4-coding-tool-eind.pdf, accessed 8/1/2019
**Abbreviations and organisation specific job titles**

In work organisations, people may use abbreviations or very specific words for their job title. As it is impossible to trace these abbreviated job titles across countries, the database does not include these very specific job titles. They do not provide a match, challenging the respondent to key in a related occupational title, as Screenshot 2 shows.

### Screenshot 2
*Screenshot showing that an abbreviation (GP) does not provide a match*


**Synonyms**

Some occupations have synonym titles. As much as possible, these titles will be included in the look-up tables, because survey respondents may search on either synonym. Example: the occupations barber, hairdresser, and hairstylist are all included in the English look-up table for the United Kingdom. In other countries, the three occupational titles would possibly be translated into one title. In these languages they are only once present in that look-up table and they are assigned the 5-digit code of the most generic term of the synonyms. Synonym occupational titles will have different 5-digit codes, but the same 4-digit code. Note that a look-up table for semantic matching cannot assign two different codes to the same occupational title. Hence, for each language the look-up table consists of unique titles only.

A very special case of synonyms are occupational titles written in reverse order, for example a *Cleaner aircraft* and an *Aircraft cleaner*. For the sake of semantic matching both ways facilitate matching. In such cases the look-up tables will therefore include one entry only.\(^{21}\)

One could consider slang words for occupations as synonyms. For example, the word wig picker is slang for psychiatrist. However, slang words are not included in the database and respondents will not be provided with a match, as Screenshot 3 shows. These respondents will understand that they need to enter another word or use the search tree. Moreover, slang words are very country specific and translation of slang words in other languages will not result in recognizable occupational titles, if these slang words can be translated at all.

### Screenshot 3
*Screenshot showing that a slang word does not provide a match*


\(^{21}\) Note that the use of reverse titles is frequently used in coding indexes, because their primary purpose is to act as a manual look-up table. These reverse titles are precisely the reason that a coding index cannot simply be used as a look-up table in a web-survey. Survey respondents will get confused if their effort results in two matches with the same meaning: which one should they tick?
Skill levels, professionals, associate professionals and certified occupations

ISCO-08 distinguishes four skill levels, referring to the ISCED educational categories (ILO 2012). An occupation’s required skill level is the basis of ISCO’s skill level structure. This challenges measurement issues that have to be solved in the source list of the database, as will be discussed here.

A first issue is that ISCO uses the words professional and associate professional to distinguish between the highly skilled and the skilled professionals. Using the semantic matching respondents will definitely not use these words to indicate their occupation. This challenges the source list to phrase occupational titles such that respondents will assess their appropriate skill level. For example, in the database the word engineer is used only for highly skilled occupations, the word technician is used for skilled occupations and the word mechanic for semi-skilled. But even if these words are replaced by more commonly understood occupational titles, how will a respondent select between a Ship engineering technician (major group 3) and a Ship engineer (major group 2), because they both provide a match on the search term ‘ship engineer’? We have considered a solution to default classify these occupations in major group 3 and add a follow-up question about the occupation’s required educational level according to the ISCED classification to decide whether the respondent should be classified in major group 2. We have, however, not applied this solution, because it would make the use of the tool more difficult.

Another solution is to include a reference to the required education. In the Netherlands, for example, the associate nurse is distinguished from the professional nurse because the occupational title includes a reference to the required education, notably ‘verpleegkundige (mbo)’ and ‘verpleegkundige (hbo)’ [nurse with medium vocational training and nurse with higher vocational training]. In Belgium, as learned from the EurOccupations expert, only one nurse occupation exists. Therefore, the look-up table for nl_BE does not include an occupation title for the Nursing associate professional. Of course, this approach is based on country-level information, which obviously was not available for all countries at stake. In these cases the database includes the translated occupational titles.

Table 2  Code, source list and two look-up tables for two occupational titles

<table>
<thead>
<tr>
<th>Code</th>
<th>B</th>
<th>BO</th>
<th>ni_BE</th>
<th>ni_NL</th>
<th>BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>22210600000000</td>
<td>Hospital nurse</td>
<td>Verpleegkundige</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32210100000000</td>
<td>Nursing associate professional</td>
<td>Verpleegkundige (hbo)</td>
<td>Verpleegkundige (mbo)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A second issue relates to the distinction between licensed or certified and non-licensed occupations, such as the Accountant, as shown in Table 3. For EU-countries, the database of licensed occupations has been used to identify the licensed occupations.
Table 3 Code, source list and two look-up tables for six occupational titles

<table>
<thead>
<tr>
<th>occupai3_API_13dgt</th>
<th>MASTER LABEL</th>
<th>en_GB</th>
<th>nl_NL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2411000200016</td>
<td>Accountant - chartered and certified</td>
<td>Accountant (chartered)</td>
<td>Registeraccountant (RA)</td>
</tr>
<tr>
<td>2411010000000</td>
<td>Accountant</td>
<td>Accountant (not-licensed)</td>
<td>Accountant, administratieconsulent (AA)</td>
</tr>
</tbody>
</table>

**Occupations in the corporate hierarchy**

Large and medium-sized organisations usually have a well-developed division of work, shaping hierarchical demarcation lines between occupations. For several reasons, issues related to the corporate hierarchy have to be solved in the database. First, ISCO-08 has assigned different skill levels to different positions within the hierarchy. Second, respondents prefer to report their position within the hierarchy. Third, valid self-identification assumes that occupational titles are clear with respect to their position in the corporate hierarchy. Several features of the corporate hierarchy are discussed here.

Many respondents prefer to indicate the hierarchy within their occupation by adding words such as senior or junior. It requires substantive empirical work to conclude for which occupations these levels are applicable and for which they are not, and how this varies across countries. Therefore, the source list does not include separate entries for junior or senior variations of an occupational title. The same applies to respondents reporting to be the 1st, 2nd or 3rd cook, pilot, or other occupational titles, and to the word ‘all-round’. In the WageIndicator web-survey respondents are offered a follow-up survey question where these categories can be ticked, if respondents desire so. Survey-holders who want to use this survey question, can send a request to the author for receiving the source list of the question and answers as well as their translations.22

Respondents also prefer to indicate when they occupy a supervisory position. ISCO-08 classifies supervisors in the same unit group as the most skilled workers supervised, while department managers are classified in Major Group 1 Managers. For survey respondents, the boundaries between the two groups may not be too clear and valid coding requires that they are perceived to be distinct. Therefore, the database includes a list of department managers (coded in Major group1) and a long list of first line supervisors (coded in Major groups 3 - 8), see Table 4.

Table 4 Examples of first line supervisors and department managers

<table>
<thead>
<tr>
<th>occupai3_API_13dgt</th>
<th>MASTER LABEL</th>
<th>occupai3_API_13dgt</th>
<th>MASTER LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>3341010000000</td>
<td>First line supervisor front-office tellers</td>
<td>1211020000000</td>
<td>Finance department manager</td>
</tr>
<tr>
<td>3341020000000</td>
<td>First line supervisor office clerks</td>
<td>1212020000000</td>
<td>Personnel department manager</td>
</tr>
<tr>
<td>3341030000000</td>
<td>First line supervisor back-office clerks</td>
<td>1219050000000</td>
<td>Administrative services department manager</td>
</tr>
<tr>
<td>3355030000000</td>
<td>First line supervisor police inspectors or detectives</td>
<td>1219980000000</td>
<td>Department manager (all other)</td>
</tr>
<tr>
<td>4222060000000</td>
<td>First line supervisor call centre agents</td>
<td>1221030000000</td>
<td>Marketing department manager</td>
</tr>
<tr>
<td>5120030000000</td>
<td>First line supervisor food preparation workers</td>
<td>1221040000000</td>
<td>Sales department manager</td>
</tr>
<tr>
<td>5120080000000</td>
<td>First line supervisor food servicing workers</td>
<td>1321010000000</td>
<td>Manufacturing department manager</td>
</tr>
<tr>
<td>5151010000000</td>
<td>First line supervisor housekeeping workers</td>
<td>1324090000000</td>
<td>Purchasing department manager</td>
</tr>
</tbody>
</table>

The concepts of careering, job ladders and job-enlargement blur the demarcation lines across occupations, whereas clarity is critical for valid self-identification. In the
EurOccupations project this turned out most problematic for the assistant occupations (Tijdens, 2010). Is the assistant plumber part of a job ladder to become a plumber and thus are the two titles one occupation, or are both separate occupations? This will vary worldwide and therefore in the source list the word assistant has been avoided as much as possible and where relevant replaced with the word helper.

**Composite jobs**

Small organisations tend to employ workers in composite jobs. Respondents may therefore want to classify themselves in more than one occupation. Web-surveys using search trees have two solutions to this problem. The first solution includes an instruction to the survey question that the occupation should be selected on which most time is spend. The second solution is allowing respondents to tick more than one occupation. Unfortunately, due to technical constraints the WageIndicator web-survey does not facilitate a second choice, but may do so in the years to come. Survey-holders can program their surveys such that the ‘what is your occupation’ question pops up twice if they want to ensure to capture the composite jobs adequately.

**Handicraft workers and machine-operators**

In the course of the 20th century, small-scale workshops have been replaced by factories and craft occupations by machine-operators due to industrialisation and technological innovations. Countries vary with respect to the degree that these processes have taken place. Nevertheless, even in highly industrialised countries traditional craft occupations continue to exist, supplying handicraft goods for commercial markets. In the source list the machine operator and the handicraft workers are assigned distinct occupational titles, e.g. the *Handicraft weaver* and the *Machine operator weaving* (see Table 5), or the *Handicraft leather worker* and the *Shoemaking machine operator*. For food manufacturing, the word handicraft worker is not applicable. For bakers and butchers, in most countries the occupational titles will refer primarily to retail trade and a different phrasing is used for comparable occupations in manufacturing.

**Table 5. Weavers and handicrafts workers**

<table>
<thead>
<tr>
<th>Code</th>
<th>Master Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>7318000600018</td>
<td>Cloth weaver</td>
</tr>
<tr>
<td>7318000900018</td>
<td>Jacquard weaver</td>
</tr>
<tr>
<td>7318001400018</td>
<td>Tapestry weaver</td>
</tr>
<tr>
<td>7318010000000</td>
<td>Textiles weaver (handicraft)</td>
</tr>
<tr>
<td>7318020000000</td>
<td>Weaver, knitter, embroiderer (handicraft)</td>
</tr>
<tr>
<td>8152010000000</td>
<td>Machine operator (weaving)</td>
</tr>
<tr>
<td>8152130000000</td>
<td>Machine operator (weaving carpets)</td>
</tr>
</tbody>
</table>

**Subsistence farmers, fishers, hunters and gatherers**

The evidence of subsistence workers has been an issue of debate. Due to rapid urbanisation worldwide, the number of subsistence workers is likely to decrease further, although an economic crisis may well lead to a temporary increase of this group. ISCO-08 includes a sub-major group, which is detailed into the *Subsistence crop farmers*, *Subsistence livestock farmers*, *Subsistence mixed crop and livestock farmers*, and *Subsistence fishers, hunters, trappers and gatherers*. The source list includes these four occupations. For the Netherlands these occupations have been translated with the extension ‘hobby’, e.g. the Subsistence livestock farmer is translated to *Veehouder* (hobby), to be distinguished from Livestock farmer (in Dutch *Veehouder*).
Male versus female titles
In some countries male and female occupational titles are different, whereas in other countries only gender-neutral or male occupational titles are in use. For example, the English word DTP operator is translated as DTP Operator/in to suit the male and the female occupational title in the German language. For the four German locales (de_AT, de_CH, de_DE, de_LU), for the two Croatian locales (hr_HR, hr_BA) and for Czech (cz_CZ) the database has these male/female occupational titles (see Table 6).

Table 6 Male and female occupational titles for Czech and German

<table>
<thead>
<tr>
<th>occupai3_API_13dgt</th>
<th>MASTER LABEL</th>
<th>cs_CZ</th>
<th>de_DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>13240500000000</td>
<td>Public transport company manager</td>
<td>Manažer(ka) veřejné hromadné přepravy</td>
<td>Leiter/in des öffentlichen Personennahverkehrs</td>
</tr>
<tr>
<td>13240600000000</td>
<td>Road, rail, water or air transport company manager</td>
<td>Manažer(ka) silniční, železniční nebo letecké přepravy</td>
<td>Leiter/in Straßen-, Bahn oder Luftverkehr</td>
</tr>
<tr>
<td>13240700000000</td>
<td>Recycling or refuse disposal manager</td>
<td>Manažer(ka) recyklace nebo spracování odpadu</td>
<td>Leiter/in Müllentsorgung oder Recycling</td>
</tr>
<tr>
<td>13240800000000</td>
<td>Energy plant manager</td>
<td>Manažer(ka) v energetice</td>
<td>Leiter/in Energieanlagen</td>
</tr>
</tbody>
</table>

Note that the source list does not distinguish between male and female titles, but in the locales the male and female titles are included (see Table 7). In a next version of the occupation database, we may add separate columns for male and for female occupational titles for the survey-holders who first have asked the respondent’s gender.

Table 7 Male and female title in English locales

<table>
<thead>
<tr>
<th>occupai3_API_13dgt</th>
<th>MASTER LABEL</th>
<th>en_AU</th>
</tr>
</thead>
<tbody>
<tr>
<td>5131000800016</td>
<td>Waiter</td>
<td>Waiter, waitress</td>
</tr>
<tr>
<td>2655000100018</td>
<td>Actor</td>
<td>Actor, actress</td>
</tr>
</tbody>
</table>

“My job title is not in your list”
Even with more than 4 000 occupational titles in a look-up table, there will be respondents who cannot identify their occupation in the semantic matching tool. This problem can be solved by making the selection of the semantic matching not mandatory. This allows respondents to key in an occupational title, not to select any of the matches and to continue the survey. For these cases, the survey-holder has to undertake office coding. As these occupational titles are likely to be rare titles, the survey-holder is advised to include a separate page with survey questions about the keyed in title, for example about the required educational level and the number of co-workers in the same occupation in the country, asked in broad classes. In case of very rare occupations it may be considered less necessary to code the occupation. This procedure also allows to detect new and emerging occupations. Depending on the sample size this office coding may be a time- and budget-consuming operation on the side of the survey-holder.

In case the survey question about occupation is mandatory, implying that the respondent cannot continue the survey unless an occupational title is selected, a survey-holder could consider adding an extra question asking whether respondent could identify his/her own occupational title and asking for a job description. This allows to identify unreliable answers, for example survey respondents with primary education selecting to be a medical doctor.

Start year and end year
Occupational structures are dynamic and the occupational composition of a labour force varies over time and across countries. Due to technological and organisational changes, new and emerging occupations arise. Most surveys ask respondents for their current occupation, but some have questions about father’s and mother’s occupation, mostly
specified as the occupation when respondent was 14 years of age. Other surveys address elderly people, who are no longer in paid employment, but are asked to report their last occupation before retirement. This issue challenges the identification of time frames for the occupational titles. A straddle carrier driver or a web manager did not exist in periods when the straddle carrier or the web was not yet introduced. Therefore, a start year can be assigned to an occupation in the source list. Hence, start years do not vary across countries. Of course, start years are only assigned to those occupations where this is known, for example in the case of web managers, where the start year 1985 is assigned.

The dynamic occupational structure also points to obsolete occupations, challenging the assignment of an end year to occupations in the look-up table. The database of historical occupations HISCO does include occupational titles that don’t exist today, because the machinery, equipment or materials used have become outdated. The HISCO scheme is based on the coding of the 1 000 most frequent male and female occupational titles in datasets from eight countries: Belgium, Britain, Canada, France, Germany, the Netherlands, Norway and Sweden. The occupational data which were employed to develop the scheme span the period 1690-1970, but are mostly from the nineteenth century. Whenever an occupation in the database is obviously outdated, an end year will be assigned. The author will assign start and end years to the titles in the source list in the years to come, as part of the SSHOC project.

---

5. The occupation to industry prediction

Developing the occupation to industry prediction\textsuperscript{24}

Many questionnaires have a question “Please write the main business activity of the organisation where you work” or similar. The answer is commonly asked as an open text field, challenging the survey-holder to code the response into an industry classification. Alternatively, in web-surveys respondents can self-identify their industry from a database. Task 8.4 in SERISS includes two deliverables, D8.10 and D8.11. For D8.10 a database of 321 industry names has been developed and translated for use in 99 countries, all coded in 3- or 4-digits, according to the classification NACE Rev. 2 (EUROSTAT, 2008).\textsuperscript{25} The database facilitates survey respondents to self-identify their industry from this lookup table by either an autosuggest box or a two-level search tree. The WageIndicator web-survey shows that respondents tend to skip the question about industry relatively more often compared to other questions, presumably because they judge answering the question as cognitively too demanding. For D8.11, an occupation>industry prediction has therefore been developed, providing survey respondents with a limited set of industries, most likely for their occupation. Of course, the limited list of industries, shown to the respondent, always includes an option ‘other’, with the full look-up table shown in the next step.

Note that the industry database, prepared for D8.10, is downloadable from the website surveycodings.org and it can be browsed in the live search tool at the website. The industry API can be used in web surveys and the industry database can be used to populate a search tree in CAPI surveys.

A multi-country occupation>industry prediction for 4-digit ISCO-08 occupations requires a dataset large enough to include as many countries as possible from among those covered by WP8. Such multi-country datasets do not exist and therefore we decided to merge datasets from several sources. We relied on the most recent waves of ESS and EWCS which use classification structures which are homogeneous and currently in place. In addition to CAPI surveys, we exploited some web-surveys, mostly the WageIndicator database. The initial idea to include controls in the predicting equations using auxiliary variables was dropped in favour of a pooled dataset NACE Rev. 2 with valid observations for only two variables: a 4-digit ISCO-08 code and a 2-digit NACE Rev. 2 code. We explored country-differences, but it turned out that the estimated most likely industries were very similar across country groups.

We then estimate a set of linear probability models (LPM) – one for each ISCO code. An LPM is a multiple linear regression model with a binary dependent variable (Wooldridge, 2010) – equal to one if the observation reported that specific ISCO unit group and 0 otherwise; the explanatory variables are given by a full set of dummy variables for the 88 divisions (i.e. 2 digits groups) included in the NACE Rev. 2 code. Estimated coefficients represent marginal effects and can be directly interpreted as a probability that each NACE division is associated with that specific ISCO group.

\textsuperscript{24} This paragraph is copied from Belloni M, Tijdens KG (2017) Occupation > industry predictions for measuring industry in surveys. Deliverable 8.11 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables

\textsuperscript{25} Tijdens KG (2016) Database of industries + explanatory note. Deliverable D8.10 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables
The occupation to industry mapping table

Based on the probability models, the WISCO database includes a sheet with a mapping to the most likely 3- or 4-digit NACE-codes for each occupational title in the database) (321 entries). Survey-holders can decide to use this mapping table for respondents’ self-identification of a 3- or 4-digit industry unit once they have selected an occupation. Across occupations the lists of 3- or 4-digits includes on average 9 entries, with a range from 2 to 33 entries. Each list ends with an industry code 9999, indicating that respondent could not self-identify with the selected list and want to select another industry. If ticked, the respondent is shown the full search tree of the industry database, as detailed in D8.10.

A next version of the WISCO database will include a mapping to the 2-digit NACE code, for survey-holders who do not aim for a greater level of detail in the industry coding.

The occupation to industry mapping table has not yet been implemented in the live search tool at surveycodings.org, but will be so before the end of the SERISS project.
6. The excel file of the occupation database

Introduction
The database has 5 sheets, which are explained hereafter.

- CODESET
- SEARCH TREE
- MAPPINGS
- LABELSET
- OCC>>INDUSTRY

Sheet CODESET
The first sheet is called CODESET (Screenshot 4). The first column shows the ISCO-08 code (first 4-digits are the ISCO code); the second column shows the source list, the remaining columns show the translations for Arabic, Indonesian, Bulgarian, Czech, Danish, etc. Note that the phrase ‘am_ET’ refers to the so-called locale, which identifies the language and the country, each using the ISO two-character identifiers, which in this case is Amharic in Ethiopia.

Sheet SEARCH TREE
The second sheet is called SEARCH TREE (Screenshot 5). It includes the structure of the search tree. The search tree consists of three levels. The first level includes 20 broad occupational groups, for which the translations are included in the sheet CODESET. The second level consists of approximately 200 detailed categories, and its translations are also included in the sheet CODESET. Note that only 2 000 of the 4 000 titles are included in the search tree, as 4 233 titles are too many to squeeze into 3 levels, and 4-level search trees are not advised in surveys. If the job cannot be found in the search tree, respondents still have the autocomplete option.
Sheet MAPPING
The sheet MAPPING includes the six variables derived from the codes used in the sheet CODESET. These variables are the four ISCO-08 codes for 1, 2, 3, and 4-digit, as well as ISCO skill level and ISEI code. When using the API in surveycodings.org, these variables are included in the dataset for each respondent.

![Screenshot 6 of the sheet MAPPING](image)

Sheet LABELSET
The sheet LABELSET has the labels (in English) of all values of the six variables included in the sheet MAPPING.

![Screenshot 7 of the sheet LABELSET](image)

Sheet OCC>>INDUSTRY
The sheet OCC>>INDUSTRY includes for all source list titles the codes of the most likely industries (see D8.10 Industry database for the codes). Note that the predictions are made

![Sheet OCC>>INDUSTRY](image)

ISEI stands for International Socio-economic Index of Occupational Status and has been developed by Harry Ganzeboom. It can be found here: [http://www.harryganzeboom.nl/isco08/index.htm](http://www.harryganzeboom.nl/isco08/index.htm), specifically the file Assignment of ISEI (International Socio-economic Index of Occupational Status) scores to ISCO-08 codes, accessed 4 Feb 2019.
for ISCO-08 4-digit, as explained in the previous section. Therefore, source titles within the same 4-digit group include the same most likely industries.

What the database cannot do
A final comment relates to the things the WageIndicator database of occupations cannot provide. The database has no mapping table to national occupational classifications, such as the classifications used in France, Germany, UK, Poland, and alike. The database is not designed for office coding. It can be used to populate tables for office coding, as for example has been done by IER U Warwick which now offers its CASCOT coding tool including several languages, based on the database. However, our database does not provide a script for removing typing errors, and unidentified titles need to be coded manually.
7. **Surveycodings.org: select country and browse the Database live search**

The website surveycodings.org allows prospective users of the database to browse the database. Screenshot 8 shows that first a country and language need to be selected, and then the autocomplete box or the search tree can be used.

**Screenshot 9**  
*Screenshot of the live search of the WISCO occupation database*

![Database live search interface](image)

*Screenshot 9 shows the list of matches for the characters 'hort'.*

**Screenshot 10**  
*Screenshot of the autocomplete box in live search of the WISCO occupation database*

![Autocomplete interface](image)

**Source list**

Screenshot 10 shows the description of the source list, and the related codes.

**Screenshot 11**  
*Screenshot of the information of the WISCO occupation database*

![Information details](image)
**Search tree**

Screenshot 11 depicts the search tree in the WISCO occupation database.

*Screenshot 12  The search tree in the WISCO occupation database*

```
Tree

- Agricultura, naturaleza, animales, medio ambiente
  - Servicios de apoyo (interno)
  - Ganado

  Horticultura, jardinería
    - Arquitecto de paisajes
    - Arquitecto de jardines
    - Florista
    - Técnico de horticultura
    - Productor hortícola
    - Jardinero ornamental
    - Jardinero de cementerio
```
References


Appendix: Instructions to translators

Introduction
The translation applies to a list of occupational titles, used in a web-survey for the question: ‘What is your occupation?’

The web visitor has to navigate the list to self-identify his/her occupation by means of semantic matching. This is similar to – for example – Google Search.

The list is drafted in English and needs to be translated in the specified languages.

Translation instructions
Please translate all occupational titles in singular (as is the case in English).

If your language uses masculine/feminine translations for occupational titles (as for example in German), please include the feminine extension.

If you have to choose between a literal translation and a translation that is used in the national labour market, please use the title reflecting the labour market reality.

If one English title can be translated with two equivalent words, which both are used frequently, the two words can be included, using a semicolon. However, if one of these words is used often and the other word is used rarely, only the often used equivalent word should be included.

If in your language words are written together (e.g. eisenbahnhinandherschreiber in German), please use blanks in these words where possible, to facilitate reader’s quick understanding.

In case two distinct English occupational titles are translated similarly in your language, this is not a problem. We will later remove similar titles.

Translation of specific groups
The word 'Manager' is used solely for people who manage a company of a department of substantial size.

The word 'Engineer' is solely used for highly skilled occupations.

The word 'Technician' is solely used for skilled occupations.

The word '(certified)' is only added for highly skilled occupations.

The word '(not certified)' is only added for skilled occupations.

The word 'farmer' is used for people who are heading a farm with - at most - few staff. In case of a large, almost industrial farm, the occupation is called 'farm manager'.

The words 'mechanic', 'machine operator', 'servicer', or 'installer' are used solely for semi-skilled occupations.

The word 'hand' – for example ‘farm hand’ - is solely used for unskilled occupations.
When you don’t know the occupation, you may use the Occupation Quick Search of ONet, see https://www.onetonline.org/find/

**Question**
For any questions, please email Kea Tijdens, University of Amsterdam, k.g.tijdens@uva.nl